Docket No.: 23623-7060

INFORMATION RICH LIBRARIES

Inventors:

Volker Schellenberger
914 Moreno Avenue
Palo Alto, CA 94303
Citizenship: Germany

Donald P. Naki
1889 Sunset Boulevard
San Diego, CA 92103
Citizenship: U.S.A.

Thomas B. Morrison
3767 Redwood Circle
Palo Alto, CA 94306
Citizenship: U.S.A.

Prepared By:

McCutchen, Doyle, Brown, & Enersen, LLP
Three Embarcadero Center, Suite 1800
San Francisco, California 94111
(650) 849-4908

Express Mail Label No. EL893834681US

INFORMATION RICH LIBRARIES

CROSS-REFERENCES TO RELATED APPLICATIONS

This application claims the benefit of priority to U.S. Provisional Patent Application No. 60/239,476, filed October 10, 2000.

STATEMENT AS TO RIGHTS TO INVENTIONS MADE UNDER FEDERALLY SPONSORED RESEARCH AND DEVELOPMENT

Not Applicable.

TECHNICAL FIELD

This invention relates to methods for producing information rich polynucleotide libraries and articles and compositions useful therein and produced thereby.

BACKGROUND OF THE INVENTION

There is currently no effective way to systematically screen all possible permutations of a polymeric biological molecule such as a polynucleotide or protein for a property of interest where the molecule is of significant length. To test four nucleotides and 20 amino acids at each position in a polynucleotide or protein, respectively, rapidly leads to a geometric increase in the number of molecules to be tested such that available methods of synthesis, and even available volumes for testing, are quickly exceeded for even a small length of such a polymer. Furthermore, even if it were physically possible to screen all permutations of a sequence of a given length, the brute force nature of such an approach would result in a great deal of the effort expended being wasted in producing and characterizing molecules lacking the desired activity.

As a compromise, a number of different approaches have arisen to sample some of the diversity available in such polymeric biological molecules.

There are two well known methods for attempting to improve the function of a protein. In random mutagenesis, one introduces random mutations and then screens for mutants with a desirable change. Although introducing more mutations per gene increases the chances of finding genes with interesting functions, each mutation potentially leads to a non-functional

protein (for instance by interfering with folding). Thus, if in creating a protein variant library, one increases the average number of mutations per gene, one then also increases the fraction of genes in the library that encode proteins which lack function.

Another method utilizes recombination between homologous coding sequences. The key advantage of recombination over random mutagenesis is that it introduces mutations known to function in a homologous protein. As a result, one generates libraries which have a relatively large diversity yet still contain a large fraction of functional mutants. In other words, recombination uses the information contained in homologous sequences to introduce diversity into a protein of interest. However, diversity in recombination is limited by the kind of information it can utilize (*i.e.*, it uses only homologous sequences) and recombination is limited in the way it utilizes that information. For example, one has limited control over the selection of crossover points. In another example, recombination usually moves regions of a gene (10-1000 bp). It rarely moves an individual residue from one sequence into a homologous position in another sequence.

Systematic approaches to altering residues in biological polymers have been made. See, for example, the "SELEX" procedures described in Tuerk et al., *Proc Natl Acad Sci U S A* 1992 Aug 1, 89(15):6988-92, and the screening for aptamers as described in Bock et al., *Nature* 1992 Feb 6, 355(6360):564-6. Pools of degenerate molecules are tested for a desired activity and the molecules possessing the greatest level of such activity can be propagated and subjected to further rounds of mutagenesis and selection. Again, however, it is not possible to test all permutations of a sequence of any significant length, so such techniques are limited by a type of "founder effect" controlled by the number of different molecules actually present in the starting population.

Systematic approaches to mutating every position in a protein have also been performed. However, the diversity at any given position is typically limited to a single change. Furthermore, such changes are typically made and assayed individually, are not made in the form of a library, and therefore do not test for multiple mutations which may be required for any given mutation to exhibit its potential activity. In some cases, a number of multiple mutants have been made at different positions throughout a protein. However, these are again typically predefined, and do not result in the production of a library of different polymers.

Thus, there remains a need in the art for a mechanism to increase the diversity of polymeric biological molecules present in a library and to increase the proportion of members of that library having a desired activity.

SUMMARY OF THE INVENTION

Methods to create information rich libraries, that is libraries that contain a high fraction of biological polymers having a desired activity are disclosed. The information used to create these libraries can include: multiple sequence alignments, substitution matrices, three dimensional structure, and prior knowledge about the structure and/or function of the reference sequence from which the library is to be produced of from a homologous sequence in a related molecule.

Generally speaking, the steps towards the manufacture of the libraries of this invention include generating a probability matrix, generating a constraint vector, designing a substitution scheme based on the probability matrix and constraint vector. The substitution scheme has utility as produced, and can be used to construct a library based thereon. The library can then be screened and the members of the library characterized. Data mining techniques can be employed to characterizing the functional clones. Optionally, the characterization data can be used as information in a subsequent iteration of the method to obtain a molecule with even more desirable properties.

Additionally, combinations of the methods described herein can be made with other techniques such as family shuffling and/or systematic scanning approaches can be performed in any order and for any number of iterations to produce the products described herein; such combinations are within the scope of the invention. Also provided are vectors containing polynucleotides produced by the disclosed methods, host cells comprising such vectors, proteins encoded by such polynucleotides, and libraries of members so generated.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a graphical representation of the relationship between a probability matrix and a constraint vector of this invention. After a probability matrix is generated, a constraint vector can be applied to the matrix to determine which amino acid substitutions will be selected to test for their effect on a desired functionality. In this graphical representation, the residues for

which values calculated by the matrix rise above the constraint put on by the vector are candidates for the library.

Figure 2 is an alignment of the sequence of ampC proteins from seven different organisms.

DETAILED DESCRIPTION OF THE INVENTION

The prior art is replete with examples of techniques intended to improve the function of proteins and polynucleotides under defined conditions. One of the most well known examples utilizes crossover recombination or DNA shuffling. Diversity produced by DNA shuffling is limited to the parent sequences and random mutations.

The invention described herein can be used to introduce residues that are not contained in the parent reference sequence but that are still likely to preserve structure and function. Because a constraint of functionality is placed on the possible mutations, the fraction of inactivating mutations is minimized. This allows one to test higher mutation frequencies and increases the chance of finding useful double and triple mutations. For example, in a library of double mutants there is one chance per member to find interacting mutations. However, if one can generate a library of members of which 100% are active and contain 20 mutations per member then there are 190 possible pair-wise interactions between these mutations per member. In addition, the library will contain a large number of functional proteins with triple and higher mutations.

DNA shuffling recombines linear blocks of sequence. This places many amino acids into new environments at the same time because residues which are close in linear sequence are not necessarily close in three dimensional space. Conversely, computer shuffling techniques allow one to recombine residues which are close in three dimensional space. Thus, one can effect mutations in subdomains of the protein which are distant in linear sequence but close in structure, thus further increasing the chance to find interacting mutations.

Because DNA shuffling recombines linear blocks of sequence, beneficial mutations at one locus may be masked by detrimental mutations nearby. For illustration purposes only, Ballinger found that recruiting a furin residue into position 104 of *Bacillus amyloliquefaciens* subtilisin improved performance of the enzyme. However, recruiting a furin residue at position 107 abolished expression of the protein. Because these residues are very close, the chances of

having a crossover event between them using DNA shuffling is remote and the resultant protein would not be active (if present at all) even though it contained a useful mutation. Ballinger, *Biochemistry* **34**:13312 (1995); Ballinger, *Biochemistry* **35**:13579 (1996).

Benefits of the invention described herein include greater control of the complexity of the library. For example, if a large number of functional proteins are desired, the constraint matrix can be constructed to include fewer substitutions likely to lead to non-functional proteins. If more diversity is desired, the constraint matrix can be constructed to provide a lower constraint on the probability matrix.

Because a library that has a higher percentage of mutated and functional proteins can be constructed, fewer members of the library are needed to achieve a suitable number of possible useful proteins. In a particular embodiment, one may characterize the sequence and function of most or all members of a population, including non-functional proteins. Thus, in addition to obtaining useful proteins with a minimal number of screening assays, one is able to obtain information as to which mutations are detrimental to a protein. This information can then be used in a new constraint matrix, for example for another iteration.

Knowledge-based approaches can incorporate information from mutation of the reference sequence into the substitution scheme. Such information can be derived from intentional mutagenesis, either sporadic or systematic, or can incorporate information from naturally occurring mutations. Systematic approaches can include saturation scans where each residue of a protein is individually changed to each of the other 19 genetically coded amino acids and the resulting single mutants screened for the desired property, as well as deletion mutagenesis scans where one or more residues are deleted from the protein, insertion mutagenesis scans where one or more residues are inserted in the protein, and alanine scanning mutagenesis where each residue of the protein is systematically replaced with an alanine. Although systematic approaches provide the most information, any mutation which provides information about the protein's ability to tolerate a mutation affecting the desired property can be used.

Before the present invention is described in detail, it is to be understood that this invention is not limited to the particular methodology, devices, solutions or apparatuses described, as such methods, devices, solutions or apparatuses can, of course, vary. It is also to

be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to limit the scope of the present invention.

Use of the singular forms "a," "an," and "the" include plural references unless the context clearly dictates otherwise. Thus, for example, reference to "a polynucleotide" includes a plurality of polynucleotides, reference to "a substrate" includes a plurality of such substrates, reference to "a variant" includes a plurality of capture probes, and the like.

Terms such as "connected," "attached," "linked," and "conjugated" are used interchangeably herein and encompass direct as well as indirect connection, attachment, linkage or conjugation unless the context clearly dictates otherwise. Where a range of values is recited, it is to be understood that each intervening integer value, and each fraction thereof, between the recited upper and lower limits of that range is also specifically disclosed, along with each subrange between such values. The upper and lower limits of any range can independently be included in or excluded from the range, and each range where either, neither or both limits are included is also encompassed within the invention. Where a value being discussed has inherent limits, for example where a component can be present at a concentration of from 0 to 100%, or where the pH of an aqueous solution can range from 1 to 14, those inherent limits are specifically disclosed. Where a value is explicitly recited, it is to be understood that values which are about the same quantity or amount as the recited value are also within the scope of the invention. Where a combination is disclosed, each subcombination of the elements of that combination is also specifically disclosed and is within the scope of the invention. Conversely, where different elements or groups of elements are individually disclosed, combinations thereof are also disclosed. Where any element of an invention is disclosed as having a plurality of alternatives, examples of that invention in which each alternative is excluded singly or in any combination with the other alternatives are also hereby disclosed; more than one element of an invention can have such exclusions, and all combinations of elements having such exclusions are hereby disclosed.

Unless defined otherwise herein, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Singleton, et al., DICTIONARY OF MICROBIOLOGY AND MOLECULAR BIOLOGY, 2D ED., John Wiley and Sons, New York (1994), and Hale & Marham, THE HARPER COLLINS DICTIONARY OF BIOLOGY, Harper Perennial, NY (1991) provide one of skill with a

general dictionary of many of the terms used in this invention. Although any methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, the preferred methods and materials are described. Unless otherwise indicated, nucleic acids are written left to right in 5' to 3' orientation; amino acid sequences are written left to right in amino to carboxy orientation, respectively. The headings provided herein are not limitations on the invention, but exemplify the various aspects of the invention. Accordingly, the terms defined immediately below are more fully defined by reference to the specification as a whole.

All publications mentioned herein are hereby incorporated by reference for the purpose of disclosing and describing the particular materials and methodologies for which the reference was cited. The publications discussed herein are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the invention is not entitled to antedate such disclosure by virtue of prior invention.

I. DEFINITIONS

The terms "polynucleotide," "oligonucleotide," "nucleic acid" and "nucleic acid molecule" are used interchangeably herein to refer to a polymeric form of nucleotides of any length, and may comprise ribonucleotides, deoxyribonucleotides, analogs thereof, or mixtures thereof. This term refers only to the primary structure of the molecule. Thus, the term includes triple-, double- and single-stranded deoxyribonucleic acid ("DNA"), as well as triple-, double- and single-stranded ribonucleic acid ("RNA"). It also includes modified, for example by alkylation, and/or by capping, and unmodified forms of the polynucleotide. More particularly, the terms "polynucleotide," "oligonucleotide," "nucleic acid" and "nucleic acid molecule" include polydeoxyribonucleotides (containing 2-deoxy-D-ribose), polyribonucleotides (containing D-ribose), including tRNA, rRNA, hRNA, and mRNA, whether spliced or unspliced, any other type of polynucleotide which is an N- or C-glycoside of a purine or pyrimidine base, and other polymers containing nonnucleotidic backbones, for example, polyamide (e.g., peptide nucleic acids ("PNAs")) and polymorpholino (commercially available from the Anti-Virals, Inc., Corvallis, Oregon, as Neugene) polymers, and other synthetic sequence-specific nucleic acid polymers providing that the polymers contain nucleobases in a configuration which allows for base pairing and base stacking, such as is found in DNA and

RNA. There is no intended distinction in length between the terms "polynucleotide," "oligonucleotide," "nucleic acid" and "nucleic acid molecule," and these terms are used interchangeably herein. These terms refer only to the primary structure of the molecule. Thus, these terms include, for example, 3'-deoxy-2',5'-DNA, oligodeoxyribonucleotide N3' P5' phosphoramidates, 2'-O-alkyl-substituted RNA, double- and single-stranded DNA, as well as double- and single-stranded RNA, and hybrids thereof including for example hybrids between DNA and RNA or between PNAs and DNA or RNA, and also include known types of modifications, for example, labels, alkylation, "caps," substitution of one or more of the nucleotides with an analog, internucleotide modifications such as, for example, those with uncharged linkages (e.g., methyl phosphonates, phosphotriesters, phosphoramidates, carbamates, etc.), with negatively charged linkages (e.g., phosphorothioates, phosphorodithioates, etc.), and with positively charged linkages (e.g., aminoalkylphosphoramidates, aminoalkylphosphotriesters), those containing pendant moieties, such as, for example, proteins (including enzymes (e.g. nucleases), toxins, antibodies, signal peptides, poly-L-lysine, etc.), those with intercalators (e.g., acridine, psoralen, etc.), those containing chelates (of, e.g., metals, radioactive metals, boron, oxidative metals, etc.), those containing alkylators, those with modified linkages (e.g., alpha anomeric nucleic acids, etc.), as well as unmodified forms of the polynucleotide or oligonucleotide.

Where the polynucleotides are to be used to express encoded proteins, nucleotides which can perform that function or which can be modified (e.g., reverse transcribed) to perform that function are used. Where the polynucleotides are to be used in a scheme which requires that a complementary strand be formed to a given polynucleotide, nucleotides are used which permit such formation.

It will be appreciated that, as used herein, the terms "nucleoside" and "nucleotide" will include those moieties which contain not only the known purine and pyrimidine bases, but also other heterocyclic bases which have been modified. Such modifications include methylated purines or pyrimidines, acylated purines or pyrimidines, or other heterocycles. Modified nucleosides or nucleotides can also include modifications on the sugar moiety, e.g., wherein one or more of the hydroxyl groups are replaced with halogen, aliphatic groups, or are functionalized as ethers, amines, or the like. The term "nucleotidic unit" is intended to encompass nucleosides and nucleotides.

Furthermore, modifications to nucleotidic units include rearranging, appending, substituting for or otherwise altering functional groups on the purine or pyrimidine base which form hydrogen bonds to a respective complementary pyrimidine or purine. The resultant modified nucleotidic unit optionally may form a base pair with other such modified nucleotidic units but not with A, T, C, G or U. Abasic sites may be incorporated which do not prevent the function of the polynucleotide. Some or all of the residues in the polynucleotide can optionally be modified in one or more ways.

Standard A-T and G-C base pairs form under conditions which allow the formation of hydrogen bonds between the N3-H and C4-oxy of thymidine and the N1 and C6-NH2, respectively, of adenosine and between the C2-oxy, N3 and C4-NH2, of cytidine and the C2-NH2, N'-H and C6-oxy, respectively, of guanosine. Thus, for example, guanosine (2-amino-6-oxy-9-β-D-ribofuranosyl-purine) may be modified to form isoguanosine (2-oxy-6-amino-9-β-D-ribofuranosyl-purine). Such modification results in a nucleoside base which will no longer effectively form a standard base pair with cytosine. However, modification of cytosine (1-β-D-ribofuranosyl-2-oxy-4-amino-pyrimidine) to form isocytosine (1-β-D-ribofuranosyl-2-amino-4-oxy-pyrimidine) results in a modified nucleotide which will not effectively base pair with guanosine but will form a base pair with isoguanosine (U.S. Pat. No. 5,681,702 to Collins et al.). Isocytosine is available from Sigma Chemical Co. (St. Louis, MO); isocytidine may be prepared by the method described by Switzer et al. (1993) Biochemistry 32:10489-10496 and references cited therein; 2'-deoxy-5-methyl-isocytidine may be prepared by the method of Tor et al. (1993) J. Am. Chem. Soc. 115:4461-4467 and references cited therein; and isoguanine nucleotides may be prepared using the method described by Switzer et al. (1993), supra, and Mantsch et al. (1993) Biochem. 14:5593-5601, or by the method described in U.S. Patent No. 5,780,610 to Collins et al. Other nonnatural base pairs may be synthesized by the method described in Piccirilli et al. (1990) Nature 343:33-37 for the synthesis of 2,6-diaminopyrimidine and its complement (1-methylpyrazolo-[4,3]pyrimidine-5,7-(4H,6H)-dione. Other such modified nucleotidic units which form unique base pairs are known, such as those described in Leach et al. (1992) J. Am. Chem. Soc. 114:3675-3683 and Switzer et al., supra.

The phrase "DNA sequence" refers to a contiguous nucleic acid sequence. The sequence can be either single stranded or double stranded, DNA or RNA, but double stranded

DNA sequences are preferable. The sequence can be an oligonucleotide of 6 to 20 nucleotides in length to a full length genomic sequence of thousands of base pairs.

A "library of DNA sequences" refers to a plurality of DNA sequences. The number of "members of the library" is not critical; it can range from less than ten to greater than 10^6. Typically in a library of DNA sequences, the library contains many different DNA sequences, all derived from the same parent DNA sequence but containing mutations in the sequence. The phrase "creating a library of DNA sequences" refers to the physical generation of a library of DNA sequences. Techniques used to physically generate a library are well know in the art and are referenced below. Typically, a "phage library" is created. "Phage libraries" comprise a DNA library incorporated into bacteriophage. The library is constructed such that the proteins encoded by the DNA library are expressed on the surface of the phage and thus on the surface of infected bacteria. The bacteria which contains the library is then "screened" for the presence of proteins with desired functionality. A "second library" is a library of DNA sequences based on the results found in the first library of DNA sequences. For example, if a beneficial mutation is found in the screening of a library, the mutation may be incorporated into the protein upon which the second library is based.

The term "IRL" refers to an information-rich library such as produced by a method of the invention.

The term "protein" refers to contiguous "amino acids" or amino acid "residues." Typically, proteins have a function. However, for purposes of this invention, proteins also encompasses polypeptides and smaller contiguous amino acid sequences that do not have a functional activity. The functional proteins of this invention include, but are not limited to, esterases, dehydrogenases, hydrolases, oxidoreductases, transferases, lyases, and ligases. Useful general classes of enzymes include, but are not limited to, proteases, cellulases, lipases, hemicellulases, laccases, amylases, glucoamylases, esterases, lactases, polygalacturonases, galactosidases, ligninases, oxidases, peroxidases, glucose isomerases and any enzyme for which closely related and less stable homologs exist. In addition to enzymes, the encoded proteins which can be used in this invention include, but are not limited to, transcription factors, antibodies, receptors, growth factors (any of the PDGFs, EGFs, FGFs, SCF, HGF, TGFs, TNFs,

insulin, IGFs, LIFs, oncostatins, and CSFs), immunomodulators, peptide hormones, cytokines, integrins, interleukins, adhesion molecules, thrombomodulatory molecules, protease inhibitors, angiostatins, defensins, cluster of differentiation antigens, interferons, chemokines, antigens including those from infectious viruses and organisms, oncogene products, thrombopoietin, erythropoietin, tissue plasminogen activator, and any other biologically active protein which is desired for use in a clinical, diagnostic or veterinary setting. All of these proteins are well defined in the literature and are so defined herein. Also included are deletion mutants of such proteins, individual domains of such proteins, fusion proteins made from such proteins, and mixtures of such proteins; particularly useful are those which have increased half-lives and/or increased activity.

"Polypeptide" and "protein" are used interchangeably herein and include a molecular chain of amino acids linked through peptide bonds. The terms do not refer to a specific length of the product. Thus, "peptides," "oligopeptides," and "proteins" are included within the definition of polypeptide. The terms include polypeptides contain co- and/or post-translational modifications of the polypeptide, for example, glycosylations, acetylations, phosphorylations, and sulphations. In addition, protein fragments, analogs (including amino acids not encoded by the genetic code, e.g. homocysteine, ornithine, D-amino acids, and creatine), natural or artificial mutants or variants or combinations thereof, fusion proteins, derivatized residues (e.g. alkylation of amine groups, acetylations or esterifications of carboxyl groups) and the like are included within the meaning of polypeptide.

"Amino acids" or "amino acid residues" may be referred to herein by either their commonly known three letter symbols or by the one-letter symbols recommended by the IUPAC-IUB Biochemical Nomenclature Commission. Nucleotides, likewise, may be referred to by their commonly accepted single-letter codes.

"Variants of a protein" are those proteins that are related to one another by a common amino acid sequence or "parental protein" but contain minor variations in amino acid sequence from each other. These changes can be conservative substitutions, non-conservative substitutions, deletions, insertions or substitutions with non-naturally occurring amino acids (mimetics). The phrase "optimizing a protein" refers to the process of changing a protein to protein variants so that the desired functionality is improved. One of skill will realize that

optimizing a protein could involve selecting a variant with lower functionality than the parental protein if that is desired.

The terms "aptamer" and "nucleic acid antibody" are used herein to refer to a single- or double-stranded polynucleotide that recognizes and binds to a desired target molecule by virtue of its shape. See, e.g., PCT Publication Nos. WO 92/14843, WO 91/19813, and WO 92/05285.

"Conservative residues" are those amino acid residues that have a similar property, such as similar chemistry. Conservative changes can be based, for example, on similar hydrophobicity, similar hydrophilicity, similar charge, similar propensity for adopting a particular secondary structure, similar shape, etc. Conservative substitution tables providing functionally similar amino acids are known in the art. In one scheme, the following six groups each contain amino acids that are conservative substitutions for one another:

1) Alanine (A), Serine (S), Threonine (T);
2) Aspartic acid (D), Glutamic acid (E);
3) Asparagine (N), Glutamine (Q);
4) Arginine (R), Lysine (K);
5) Isoleucine (I), Leucine (L), Methionine (M), Valine (V); and
6) Phenylalanine (F), Tyrosine (Y), Tryptophan (W).

(*see, e.g.*, Creighton, *Proteins* (1984)).

"Amino acid mutations" are substitutions, deletions or insertions in amino acid sequences. For example, if an alanine occurs in an amino acid sequence, the alanine could be substituted to a serine, it could be deleted or another amino acid residue could be inserted on the amino or carboxy side of the residue. Because alanine and serine are members of the same conserved family of amino acids in the scheme described above, such a substitution can be termed a "conservative substitution." Other schemes can be used.

The term "antibody" as used herein includes antibodies obtained from both polyclonal and monoclonal preparations, as well as: hybrid (chimeric) antibody molecules (see, for example, Winter et al. (1991) *Nature* 349:293-299; and U.S. Patent No. 4,816,567); F(ab')2 and F(ab) fragments; Fv molecules (noncovalent heterodimers, see, for example, Inbar et al. (1972) *Proc Natl Acad Sci USA* 69:2659-2662; and Ehrlich et al. (1980) *Biochem* 19:4091-4096);

single-chain Fv molecules (sFv) (see, for example, Huston et al. (1988) *Proc Natl Acad Sci USA* 85:5879-5883); dimeric and trimeric antibody fragment constructs; minibodies (see, e.g., Pack et al. (1992) *Biochem* 31:1579-1584; Cumber et al. (1992) *J Immunology* 149B:120-126); humanized antibody molecules (see, for example, Riechmann et al. (1988) *Nature* 332:323-327; Verhoeyan et al. (1988) *Science* 239:1534-1536; and U.K. Patent Publication No. GB 2,276,169, published 21 September 1994); and, any functional fragments obtained from such molecules, wherein such fragments retain specific-binding properties of the parent antibody molecule.

As used herein, the term "monoclonal antibody" refers to an antibody composition having a homogeneous antibody population. The term is not limited regarding the species or source of the antibody, nor is it intended to be limited by the manner in which it is made. Thus, the term encompasses antibodies obtained from murine hybridomas, as well as human monoclonal antibodies obtained using human hybridomas or from murine hybridomas made from mice expression human immunoglobulin chain genes or portions thereof. See, e.g., Cote, et al. *Monoclonal Antibodies and Cancer Therapy*, Alan R. Liss, 1985, p. 77.

The term "sequence alignment" refers to the result when at least two amino acid sequences are compared for maximum correspondence, as measured using one of the following "sequence comparison algorithms." Optimal alignment of sequences for comparison can be conducted by any technique known or developed in the art, and the invention is not intended to be limited in the alignment technique used. Exemplary alignment methods include the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* **2**:482 (1981), the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* **48**:443 (1970), the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* **85**:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), and by inspection.

The "three dimensional structure" of a protein is also termed the "tertiary structure" or the structure of the protein in three dimensional space. Typically the three dimensional structure of a protein is determined through X-ray crystallography and the coordinates of the atoms of the amino acids determined. The coordinates are then converted through an algorithm into a visual representation of the protein in three dimensional space. From this model, the local "environment" of each residue can be determined and the "solvent

accessibility" or exposure of a residue to the extraprotein space can be determined. In addition, the "proximity of a residue to a site of functionality" or active site and more specifically, the "distance of the α or β carbons of the residue to the site of functionality" can be determined. (For glycine residues, which lack a β carbon, the α carbon can be substituted.) Also from the three dimensional structure of a protein, the residues that "contact with residues of interest" can be determined. These would be residues that are close in three dimensional space and would be expected to form bonds or interactions with the residues of interest. And because of the electron interactions across bonds, residues that contact residues in contact with residues of interest can be investigated for possible mutability. Additionally, molecular modeling can be used to determine the structure, and can be based on a homologous structure or *ab initio*. Energy minimization techniques can also be employed.

Although not dependent on three dimensional space, the "residue chemistry" of each amino acid is influenced by its position in a protein. "Residue chemistry" refers to characteristics that a residue possesses in the context of a protein or by itself. These characteristics include, but are not limited to, polarity, hydrophobicity, net charge, molecular weight, propensity to form a particular secondary structure, and space filling size.

The phrase "probability matrix" refers to a matrix for determining the probability that an amino acid can be substituted with another amino acid. Typically this matrix is in the form of an algorithm that determines the probability of substitution from the amino acid and its position. The individual entries in the matrix give a probability for placing a given amino acid in the preselected reference sequence at that position. The algorithm can be based on maintenance of structure, evolutionary diversity amongst a family of proteins and/or other factors described herein, as well as combinations thereof. The phrase "generating a probability matrix" refers to the process of determining the variable upon which the probability matrix will be based and, if needed, developing the algorithm to determine the substitutions in the matrix. The probability matrix can be "normalized" by setting the probability of a particular substitution in the matrix to "1" and correspondingly adjusting the relative probabilities of the other amino acids. The matrix can be normalized to the substitution most favored at that position by the algorithm, or to the value in the matrix for the wild type residue in the reference sequence at that

position, or in any other desired manner. Normalization can be desirable to increase the degree to which mutations at a given position are sampled in generating the library.

The phrase "constraint vector" refers to a constraint put on or "applied to" the probability matrix to determine whether and the degree to which mutations at a given position in the matrix are to be included in the library. It too is typically an algorithm that determines whether a particular mutation will result in a functional protein. Variables that can be used to determine the constraint vector are also described below.

II. PROBABILITY MATRIX

A probability matrix is generated to provide an estimate that a given residue will provide a desired activity in a biological polymer of interest. The biological polymer can be a polynucleotide having its own activity of interest, or can encode a protein having an activity of interest. Biological polymers can include polynucleotides exhibiting catalytic activity, for example ribozymes, polynucleotides exhibiting binding activity, for example aptamers, polynucleotides exhibiting promoter activity, or polynucleotides exhibiting any other desired activity, alone or in combination with any other molecule.

The matrix comprises rows representing a given position in the biological polymer of interest, and columns for a plurality of different residues which can be incorporated into the reference sequence. The matrix entries give an estimate for the probability that incorporation of the residue in that column at the position in that row will produce a polymer having the desired activity.

A probability matrix can be generated for at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90 or 100 positions in the reference sequence up to the entire sequence, and can include contiguous residues or noncontiguous residues or mixtures thereof. The matrix can include at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45 or 50 different residues. Naturally occurring residues can be included in the matrix, as well as unnatural residues for synthetic methods, and combinations thereof.

A profile can be created from the matrix based on probability scores and weighting factors. The probability matrix for a protein is preferably an n x 20 matrix that calculates the probability for any point mutation of the target gene that the mutation will result in a protein having the desired function.

In one aspect, a probability matrix is calculated for a given protein library to be produced. To do this, numerical values are assigned to each amino acid that can be substituted into the sequence. One of skill will realize these numbers are arbitrary in that they are relative to each other only for the particular library being produced. It can be useful in some instances to assign the wild type residue at a given position a value of 1, although the wild type residue can be assigned any value. From this initial value, the values of each of the 20 encoded naturally occurring amino acids at each position can be assigned.

In some instances, it can be useful to assume, initially, that the wild type residue is a useful residue and results in a functional molecule. Thus, the value of most other residues should be less than that given to the wild type, therefore in the present example, less than "1". Furthermore, in assigning values, residues that exhibit a low degree of conservation in homologs can be given large values in the probability matrix. Also, because areas of a protein which allow an insertion should be more tolerant to substitution, higher probabilities can be given to nonnative residues at positions which are close to insertions or deletions in homologs.

An example of a ranking of amino acid for valuation in this invention can be found in Gribskov, *Proc Nat'l Acad* Sci *USA* **84**:4355 (1987). The degree of conservation for each position can be used to scale the values according to Gribskov.

Other information can be used to generate a probability matrix. For example, structural information has been found to be useful. As is well known, Hidden Markov models calculate the probability of going from one residue to the next based on sequence alignments. These models also include probabilities for gaps and insertions. See, Krogh, "An introduction to Hidden Markov models for biological sequences," in COMPUTATIONAL METHODS IN MOLECULAR BIOLOGY, Salzberg, et al., eds, Elsevier, Amsterdam.

Other structural information found to be useful is the three dimensional structure of the protein. See for example, Dahiyat & Mayo, *Protein Sci.* 5:895 (1996). This can be determined crystallographically or from molecular modeling techniques. Energy minimization methods can also be employed.

A variety of different substitution matrices can be used as input for the calculation of a probability matrix. The choice of substitution matrix will impact the probability and ultimately the mutagenesis scheme. Thus, if mutations based on sequence alignment are desired, a sequence alignment substitution matrix should be chosen. Alternatively, if mutations that depend on general mutability are desired, a substitution matrix reflecting this need should be chosen.

Substitution matrices can be calculated based on the environment of a residue, *e.g.*, inside or accessible, in α-helix or in β sheet. See, Overington, et al., *Protein Sci* 1:216 (1992). Methods to determine solvent accessible residues are known in the art. *See*, for example, Hubbard, *Protein Eng* **1**:159 (1987).

More complex substitution matrices which consider secondary structure, solvent accessibility, and the residue chemistry are also suitable for use in probability matrices. *See*, for example, Bowie & Eisenberg, *Nature* **356**:83 (1992).

One of skill will realize that a probability matrix can require quite complex mathematical calculations and therefore an algorithm that determines the matrix can be desired or even required. The development of such an algorithm is within the skill in the art following the teachings herein. Similarly, because of the complex calculations necessary to carry out the algorithm, it can be desirable to generate a computer program and employ it on a computer to calculate the probability matrix. Again, this is within the skill in the art.

III. CONSTRAINT VECTORS

The constraint vector preferably should reflect the likelihood that a specific mutation at each amino acid position of a protein will improve or affect the desired function of that protein. One example of a constraint vector is a correlation matrix. The constraint vector can also include knowledge-based component(s), such as prior knowledge of effects of single

mutations, for example from mutagenesis scans or from naturally occurring mutations which affect the function of interest.

Another example is based on proximity. For example, it can be assumed that residues which are close to the active site of an enzyme are more likely to affect enzyme activity and or specificity than more distant residues and thus, a mutation of a residue near the active site will affect the activity and/or specificity (either positively or negatively) than a mutation further away from the active site. The same proximity argument can be used for other applications: proximity to an epitope, proximity to an area of structural conflict, proximity to a conserved sequence, proximity to a binding site, proximity to a cleft in the protein, proximity to a modification site, etc.

There are a variety of methods available to estimate distance, and any technique known or developed in the art for estimating such distances can be used. For instance, the library can be constrained by distance of α or ß-carbons to the active site of an enzyme. In another embodiment, the constraint can be based on the residues that make contact with the residues of interest (=1^{st} shell) and residues which contact the residues in the 1^{st} shell (= 2^{nd} shell).

In another example, the simple distance function between β carbons of the enzyme and the β carbon of a bound ligand can be used to constrain a library. A linear function can be used where the threshold of acceptable mutations depends on the distance from the bound ligand. However, one can also utilize a variety of other functional relationships between distance and threshold of mutability, *e.g.*, the square of the distance or the square root of the distance.

The physical distances from a known crystal structure of the reference sequence can be used. Alternatively, molecular modeling approaches can be used. For example, the structure of the reference sequence can be predicted based on its homology to a known structure, and then used to calculate distances. Or the entire structure of the reference sequence can be predicted and distances then calculated from the predicted structure. Energy minimization methods can be used.

Another way to generate constraint vectors is through correlation in evolutionary data. It has been observed that the replacement of a residue in a protein or protein family can be correlated with replacements in other positions. See, Lockless & Ranganathan, *Science* **286**:295 (1999); and Gobel, et al., *Proteins* **18**:309 (1994). In such cases it may be advantageous to design the constraint vector such that all correlated residues are mutated simultaneously.

Conservation Indexes can be used as the elements of a constraint vector. In this capacity, one can avoid mutating residues that are highly conserved, or conversely, focus mutations on conserved regions of the protein. Algorithms for calculating Conservation Indexes at each position in a multiple sequence alignment are known in the art (Novere et al. Biophys. Journal v.76 , p. 2329-2345, May 1999).

One of skill will realize that, like a probability matrix, generation of a constraint vector can require quite complex mathematical calculations and therefore an algorithm that determines the vector may be desired or even needed. The development of such an algorithm is within the skill in the art following the teachings herein. Similarly, because of the complex calculations necessary to carry out the algorithm, it can be desirable to generate a computer program and to employ it on a computer to generate the constraint vector. Again, this is within the skill in the art following the teachings herein.

IV. APPLICATION OF THE CONSTRAINT VECTOR TO THE PROBABLILITY MATRIX TO PRODUCE A SUBSTITUTION SCHEME

To determine which positions are to be permuted and which new residues will be tried in those positions, the constraint vector is applied to the probability matrix. This is done to increase the chance of finding improved variants and to decrease the risk of producing mutants with undesired properties, while generating a library of a size which can be effectively screened for a desired property. This application can also determine the degree to which a given change will be represented in the library, or a simpler threshold approach can be used, wherein all changes at a given position which meet the criteria imposed by the constraint vector are equally represented in the library.

An exemplary algorithm is shown in Figure 1. As is graphically represented in Figure 1, the constraint vector can be imagined as being "lowered" onto the probability matrix.

Positions in the probability matrix which are higher than the corresponding value in the constraint vector (i.e., which exceed the threshold imposed by the constraint vector) are candidates for mutagenesis. As the constraint vector is lowered, the number of positions to be mutagenized increases, and the number of new substitutions at each position increases. The degree to which the constraint vector is lowered is thus a determining factor in the size of the library which results. Application of the constraint vector can thus itself be constrained by the desired size of the library; a predetermined library size can be used to determine the degree to which the constraint vector allows the probability matrix to be sampled.

The substitution scheme produced by applying the constraint vector to the probability matrix is itself a useful result. The substitution scheme can be provided and used to create a library. The substitution scheme can be subjected to additional constraints prior to being employed in creating a library. For example, knowledge-based approaches can incorporate information about the activity of the polymer of interest and can be used to focus the substitution scheme to identify residues more likely to result in the desired activity when substituted as well as in identifying residues less likely to result in the desired activity.

One of skill will realize that the application of a constraint vector to a probability matrix can require quite complex mathematical calculations and therefore an algorithm that applies these two algorithms may be desired or required. The development of such an algorithm is within the skill in the art following the teachings herein. Similarly, because of the complex calculations necessary to carry out the application algorithm, it can be desirable to generate a computer program and employ it on a computer to do this. Again, this is within the skill in the art following the teachings herein.

V. CONSTRUCTION OF A LIBRARY

The simplest randomization scheme for polynucleotides encoding proteins is codon-based mutagenesis. In other words, after the amino acid residues to be mutated have been identified, the corresponding codons in the corresponding DNA sequence are randomized to create a DNA library. Procedures to randomize codons are known in the art (Huse et al., Int Rev Immunol. 1993;10(2-3):129-37; Kirkham et al., J Mol Biol. 1999 Jan 22;285(3):909-15). As

one of skill will appreciate, more complicated randomization schemes can be designed which are more compatible with nucleotide-based mutagenesis.

Codon mutagenesis can be done in equimolar ratios, *e.g.*, for a given site all mutagenic oligomers are added in equimolar ratios, or in ratios that relate to the probability matrix and/or the constraint vector. For example, one can bias a library in favor of mutations which are more likely to result in a functional protein. If desired, wild type oligos can be added to adjust the overall frequency of mutagenesis for a position or a region of the target gene.

In one embodiment, nucleotide-based randomization is used. This method has two advantages over synthesizing individual oligos for each substitution: it is less expensive as fewer oligos are needed; and the library will contain clones where neighboring (in linear sequence) positions have been simultaneously mutated.

Nucleotide-based mutagenesis can be optimized to produce a desired set of amino acids (Goldman & Youvan, *Bio/Technology* **10**:1557 (1992); Huang & Santi, *Anal Biochem* **218**:454 (1994); Jensen, et al., *Nucleic Acids Res* **26**:697 (1998); and Tomandl, et al., *J. Comp.-Aided Molec. Design* **11**: 29 (1997)). These authors did not consider a probability matrix; their focus was on inclusion of a desired set of amino acids. Nucleotide mixtures which encode amino acids mixtures that optimally conform to the calculated probability matrix and constraint vector can be calculated and synthesized.

Alternatively, portions of a coding region or an entire coding region can be chemically synthesized in a codon-by-codon technique using mixtures of activated trinucleotides at the positions to be substituted. In this way, only the desired codons are incorporated, dysfunctional mutations inevitably resulting from nucleotide-based randomization are avoided, and mixtures of adjacent changes can be readily provided. Additionally, controlling the degree of incorporation of a given mutation at a given position can be readily accomplished by varying the amount of the particular activated trinucleotides in the mixture for that position.

Oligonucleotide-driven site-directed mutagenesis can also be used. Suitable site-directed techniques include those in which a template strand is used to prime the synthesis of a complementary strand lacking a modification in the parent strand, such as methylation or

incorporation of uracil residues; introduction of the resulting hybrid molecules into a suitable host strain results in degradation of the template strand and replication of the desired mutated strand. *See* Kunkel, Proc Natl Acad Sci U S A 1985 Jan;82(2):488-92; QuikChange™ kits available from Stratagene, Inc., La Jolla, CA. Mixtures of individual primers for the substitutions to be introduced can be simultaneously employed in a single reaction to produce the desired combinations of mutations. Simultaneous mutation of adjacent residues can be accomplished by preparing a plurality of oligonucleotides representing the desired combinations. PCR methods for introducing site-directed changes can also be employed.

Oligos synthesized from mixtures of nucleotides can be used. The synthesis of oligonucleotide libraries is well known in the art. In one alternative, degenerate oligos from trinucleotides can be used (Gaytan, et al., *Chem Biol* **5**:519 (1998); Lyttle, et al., *Biotechniques* **19**:274 (1995); Virnekas, et al., *Nucl. Acids Res* **22**:5600 (1994); Sondek & Shortle *Proc. Nat'l Acad. Sci. USA* **89**:3581 (1992)). In another alternative, degenerate oligos can be synthesized by resin splitting (Lahr, et al., *Proc. Nat'l Acad. Sci. USA* **96**:14860 (1999); Chatellier, et al., *Anal. Biochem.* **229**:282 (1995); and Haaparanta & Huse, *Mol Divers* **1**:39 (1995))

After the oligos which incorporate desired protein mutations are constructed, they can be assembled with the DNA that encodes the desired protein. Site-directed mutagenesis using a single stranded DNA template and mutagenic oligos is well known in the art (Ling & Robinson, *Anal Biochem* **254**:157 (1997)). It has also been shown that several oligos can be incorporated at the same time using these methods (Zoller, *Curr Opin Biotechnol* **3**: 348 (1992)). Single stranded DNA templates are synthesized by degrading double stranded DNA (Strandase™ by Novagen). The resulting product after strain digestion can be heated and then directly used for sequencing. Alternatively, the template can be constructed as a phagemid or M13 vector. Other techniques of incorporating mutations into DNA are known and can be found in, *e.g.*, Deng, et al., *Anal Biochem* **200**:81 (1992)). In an alternative embodiment, sequences are assembled by PCR fusion from synthetic oligos (Horton, et al., *Gene* **77**:61 (1989); Shi, et al., *PCR Methods Appl.* **3**:46 (1993); and Cao, *Technique* **2**:109 (1990)). PCR with a mixture of mutagenic oligos can be used to create the DNA sequences that reflect the diversity of the library.

Cassette mutagenesis can also be used in site-directed random mutagenesis. Using this technique, a library can be generated by ligating fragments obtained by oligosynthesis, PCR or combinations thereof. Segments for ligation can, for example, be generated by PCR and subsequent digestion with type II restriction enzymes. This enables introduction of mutations via the PCR primers. Furthermore, type II restriction enzymes generate non-palindromic cohesive ends which significantly reduce the likelihood of ligating fragments in the wrong order. Techniques for ligating many fragments can be found in Berger, et al., *Anal Biochem* **214**:571 (1993); and U.S. Pat. App. Ser. No. 09/566,645, filed May 8, 2000.

A problem encountered in random mutagenesis is the manufacture of stop codons at the site of diversity. *In vitro* translation can be used to obtain libraries that are free of stop codons or other artifacts (Cho, et al., *J Mol Biol* **297**:309 (2000)).

The particular chemical and/or molecular biological methods used to construct the library are not critical; any method(s) which provide the desired library can be used. For example, oligonucleotides can be inserted into a phage vector so that the phage particle expresses the encoded protein on its surface. Alternatively, one can manufacture a protein array wherein the encoded proteins are immobilized on a suitable surface and functional activity is assessed and the corresponding protein identified. In yet another embodiment, if the ability of a protein to bind to a target is the desired function, a mixture of proteins encoded by the library can be contacted with the desired target and the proteins bound identified and sequenced. For construction of libraries see, US Patent Nos. 6,114,149; 6,107,059; 5,922,545; 5,830,721; 5,723,323; 5,698,426; 5,571,698; 5,565,332; and PCT Patent Application WO 0046344.

VI. CHARACTERIZING THE LIBRARY MEMBERS

After a library is generated, the members can be characterized and the library screened for members that exhibit the desired activity. In addition to finding the desired functional protein, the information from the screen can be used to design improved probability matrix and constraint vectors for a next iteration of mutagenesis and library construction. For example, the probability matrix can be improved by determining the mutations in the gene that are compatible with expression, folding, and/or stability. Identifying stabilizing mutations or combinations of mutations can be of particular importance if library size is very limited by

expense or difficulties in cloning. Under these conditions it can be advantageous to sequence all or most clones in a library. In a subsequent round of evolution the deleterious mutations identified in the prior round can then be avoided altogether. In addition, all of the sequences present in the library can be sequenced if the number of clones to be assayed is small. It can be cost efficient to sequence even clones which have no activity because they help to improve the probability matrix. Sequencing using DNA or RNA arrays (Hyseq, Inc.) can be used.

After screening for a particular function, it can be determined which mutations affect that function. This information would help to understand the underlying mechanism of the functional protein. Furthermore, the next round of library construction can be focused on these positions and neighboring residues which produce the desired activity (*i.e.*, the constraint vector can be modified to better ensure functional proteins). The constraint vector can also be improved by determining the combinations of mutations that occur simultaneously in improved clones. These residues may interact and should be mutated simultaneously in subsequent rounds. Such synergistic mutations can be particularly important because they are almost impossible to identify by simple random mutagenesis.

Analysis of the library can also reveal the mutations that are missing from the unselected libraries. This could indicate toxicity, in addition to technical problems with library construction. If it is determined that an individual clone is toxic, such a polynucleotide or its encoded protein may find use as a drug or compound in which toxicity to bacteria is desired (assuming the library is constructed in *E. coli*). A related issue is the fitness distribution in the library. This can indicate the optimum mutation frequency for the library. The fitness distribution can also be used to compare various methods of calculating the probability matrix and the constraint vector, *i.e.*, the presence of continuous improvements of these methods.

Other useful products produced by the method of the invention include polynucleotides incorporating mutations identified through construction and screening of such libraries, vectors (including expression vectors) comprising such polynucleotides, host cells comprising such polynucleotides and/or vectors, and libraries of biological polymers, and libraries of host cells comprising and/or expressing such libraries of biological polymers.

VII. CORRELATION BETWEEN STRUCTURE AND FUNCTION OF PROTEIN MUTANTS

Statistical analyses of the correlation between structures and functions of molecules have been widely used to guide the optimization of small molecule drugs (quantitative structure activity relationship, or QSAR). One can differentiate between parameter-free approaches (for example Free, *J. Med. Chem.* (1964)) and methods which consider various physico-chemical parameters of the various substituents of a molecule (for example Carotti, *Chem Biol Interact* **67**:171 (1988)). See also, Goldman, et al., *Drug Development Research* **33**:125(1994) and Lahr, et al., *Proc. Nat'l Acad. Sci. USA* 96:14860 (1999). Either approach can be used for the libraries of the instant invention. In addition one can use algorithms based on the 3D structure of the protein of interest.

The amino acid sequence can be determined for variants that exhibit desired properties. The variants may each contain multiple mutations with respect to the parent molecule, and several variants may share one or more identical mutations while having other, nonshared mutations. The data mining task is to assign the degree to which individual mutations or combinations of mutations contribute to the observed improvement in properties, and to identify which pairs or groups of amino acids interact with each other (i.e. the observed measured property for the combined mutations is non-additive compared to the effect of the mutations individually). Methods for performing this data mining are known in the art; computer programs implementing suitable techniques are available (e.g., Spotfire).

VIII. CO-VARIATION AS A TOOL TO SELECT THE REGION TO BE MUTAGENIZED

Co-variation is the tendency of some residues to change simultaneously with other residues, *i.e.*, the residues are linked during evolution. These co-variant residues can be linked by structure and/or they may be linked by function. Once coupled residues have been identified, if one of the residues is found to be a candidate for mutation, the other residue can be assigned a higher probability of being a candidate as well. In this way, mutations which otherwise would not be obvious in a probability matrix or a constraint vector can be included. For further discussion of co-variation, see Gobel, et al., *Proteins* **18**:309 (1994); Jespers, et al., *J. Mol. Biol.* **290**:471 (1999); and Pazos, et al., *Comput. Appl. Biosci.* **13**:319 (1997).

VII. UTILITY OF THE LIBRARIES OF THIS INVENTION

While the utility of the libraries of this invention will be evident to one of skill in the art, the libraries will be particularly useful in preparation of enzymes or ligands with increased activity, enzymes or ligands with modified activity, proteins with increased stability, removal of immunogenic epitopes from useful proteins, improving expression levels of proteins, and improving grafting of domains or loops into proteins.

EXAMPLES

The following examples are set forth so as to provide those of ordinary skill in the art with a complete description of how to make and use the present invention, and are not intended to limit the scope of what is regarded as the invention. Efforts have been made to ensure accuracy with respect to numbers used (e.g., amounts, temperature, etc.) but some experimental error and deviation should be accounted for. Unless otherwise indicated, parts are parts by weight, temperature is degree centigrade and pressure is at or near atmospheric, and all materials are commercially available.

Example 1: Subtilisin With Novel Substrate Specificity

GG36 (savinase) is a subtilisin protease from *Bacillus lentus*. The goal of this Example is to generate mutants of the protease that possess a novel substrate specificity.

A published multiple sequence alignment of 124 subtilisin-like serine proteases (Siezen, et al., *Protein Science* **6**:501 (1997)) was recreated from a publicly available database (GENBANK), with the sequence labeled baalkp in the database being substituted with that of GG36. GG36 differs from baalkp by only one residue substitution. In baalkp, residue 87 is an asparagine while in GG36 a serine residue is found at the corresponding position. The GG36 amino acid sequence was used as the reference sequence, and those positions of the alignment for which the GG36 sequence had a gap character were deleted.

A profile for the alignment was generated using the method of Gribskov (Gribskov, *Proc. Nat'l Acad. Sci. USA* **84**:4355 (1987)) except that a mutation probability matrix was used in place of the log-odds matrix used by Gribskov. *See* Table 1. The mutation probability matrix gives the probabilities that a given amino acid will mutate to any another amino acid in a given evolutionary interval (Dayhoff, et al., *Atlas of Protein Sequence and Structure* (Natl. Biomed.

Res. Found., Washington), Vol. 5, Suppl. 3, pp. 345-358 (1978)). The mutation probability matrix PAM 128, generated from the PAM1 matrix as described by Dayhoff, was used.

In the Gribskov method, the value of the profile for amino acid a at position p is given by

$$M(p,a) = \sum_{b=1}^{20} W(p,b) \times Y(a,b)$$

where Y(a,b) is the probability obtained from Dayhoff's mutation probability matrix for the substitution of a for b, and W(p,b) is a weight for amino acid b at position p.

The frequency of an amino acid in the alignment at a particular position was used for its weight:

$$W(b,p) = n(b,p) / N_r ,$$

where $n(b,p)$ is the number of times b appears at position p, and N_r is the total number of amino acid counts at that position.

The probability matrix, GG36 residues against all 20 substitution residues, was normalized to the largest fraction in each row. *See* Table 2.

The constraint vector was designed such that mutagenesis would focus on positions which are close to the active site of the enzyme. The calculation was based on two crystal structures which have peptides bound to different regions of the active site: a structure of FN2 (a subtilisin mutant from *B. lentus*, which is identical to GG36 except for the following substitutions; K27R, V104Y, N123S, and T174A) which contained the peptide Ala-Ala-Pro-Phe bound to the S_4 to S_1 subsites; and a structure of subtilisin BPN' (from *B. amyloliquefaciens*) which had the inhibitor Suc-Ala-Phe-Ala bound to the S'_1 to S'_3 subsites. Both structures were aligned using the program "insight II" (MSI, San Diego, CA). Subsequently, the coordinates of the inhibitor Suc-Ala-Phe-Pro-Ala were moved into the structure of FN2. The combined coordinates were imported into Excel (Microsoft, Redmond, WA). For each residue of the enzyme the distance between the beta carbon atom and the closest beta carbon atom of the two bound peptides was calculated. Where glycine residues, which do not have a beta carbon, occurred, the distance between the alpha carbon of the glycine residue and the beta carbon of the bound peptide was calculated instead.

For each backbone residue, a selection value was calculated using the constraint vector as described below. This value was used to select residues from the sequence profile for

inclusion in the substitution table. Profile values greater than or equal to the selection value were added to the substitution list for that position. The lower the value, the increased chance that a substitute residue was selected at that position.

A linear constraint vector of the formula y = mx+b was used to generate the combinatorial selection scheme, where x=Cβmin. The m and b terms were chosen to provide ~100 substitutions from residues between 1 and 10 Å from the active site as described, yielding m=0.15500 and b=-0.40000. Any y values <1 (which result from a distance of >10 Å) were ignored. Entries in the profile shown in Table 1 which exceeded the y value determined for that position by applying the constraint vector (and =1) were selected for inclusion in the combinatorial library. Application of the constraint vector to the probability matrix in this manner produced the substitution table shown in Table 3, containing 105 suggested substitutions.

Visual inspection of the enzyme structure determined that most residues which are close to the bound ligand were included in the mutagenesis scheme. It was decided to avoid mutation of positions H62 and S215 as proposed by the algorithm because these two residues are part of the catalytic triad of subtilisin. Furthermore, V66C was eliminated from the mutagenesis scheme because an unpaired Cys residue is unlikely to lead to a functional GG36. These alterations represent contribution of a knowledge-based constraint to the results produced by applying the constraint vector to the probability matrix. As the consensus sequence derived from alignment of the large family was quite different from that of savinase, the most prevalent residue at several positions in the profile was not the residue in the savinase backbone. Additionally, in some cases the wild type residue was suggested to be substituted with itself. In cases where only a single substitution of a residue was suggested, the technique used to form the library could be doped with the wild type residue to prevent inclusion of a possibly debilitating residue in all members of the library.

Example 2. Alteration of β-lactamase Specificity Using a Scoring Profile

This example demonstrates the application of a distance-based constraint vector to a position-specific scoring matrix generated using a multiple sequence alignment of seven members of the ampC family of proteins and a PAM32 substitution matrix.

To create the IRL produced in this example, 7 beta lactamase ampC protein sequences (those from *A. sobria*, *E. coli*, *O. anthropi*, *P. aeroginosa*, *S. enteriditis* and *Y. enterocolitica*) were aligned using the default parameters of the program AlignX (a component of Vector NTI Suite 6.0 from Informax, Inc.), which is an implementation of the ClustalW alignment algorithm [Thompson, J. D., D. G. Higgins, et al. (1994). Nucleic Acids Res 22(22): 4673-80.]. *See* Figure 2. The sections of the alignment for which the reference sequence (*E. cloaceae*) had a gap character were discarded, as only positions at which the reference sequence contained an amino acid were used.

The multiple sequence alignment of ampC was used to generate a profile using the method of Gribskov as described above except that a mutation probability matrix was used instead of the log-odds substitution matrix form used by Gribskov. The mutation probability matrix gives the probabilities that any given amino acid will mutate to each of the other amino acids in a given evolutionary interval. The mutation probability matrix PAM 32, which was generated from the PAM1 matrix as described [Dayhoff, M. et al. (1978) *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Washington), Vol. 5, Suppl. 3, pp. 345-358)], was used.

A distance-based constraint was applied to the scoring matrix to limit mutations to residues that are surface exposed and within 6 angstroms from the binding site of ligands in the *E. cloacae* ampC 3D structure. Specifically, the *E. cloacae* ampC crystal structure (Protein Database Base ID# 1BLS) and 6 *E. coli* ampC structures containing bound inhibitors or substrates (Protein Database Base structures 1C3B, 1FCM, 1FCN, 1FCO, 1FSW, 1FSY) were first loaded into the program MOE 2000.01 (Chemical Computing Group, Inc., Montreal Canada). Because each structure consists of a homodimer, one of the monomers and its associated ligand was deleted. Next, the main chains of all the structures containing bound ligands were aligned (0.4 angstroms RMS deviation) and all the water molecules were manually deleted. The main chains of all structures except the *E. cloacae* structure (1BLS) were then removed. The resulting structure consisted of the *E. cloacae* ampC molecule with all of the superimposed ligands from the other 6 ampC structures. All surface-exposed side chains (i.e. the beta carbon and additional atoms not in the backbone) in ampC with atoms within 6 angstroms of the ligand atoms were then selected for the IRL library. Five of the top

substitutions based on the scoring matrix were chosen at each of these sites. This library was termed the 'profile library' or IRL1 library.

To create the IRL1 DNA library, 90 mutagenic forward primers containing the different substitutions were designed and used in a PCR reaction containing a single wild type reverse primer and the *E. cloacae ampC*-containing plasmid pAL20 as template. After digestion of the methylated template DNA using the *DpnI* enzyme, the PCR product was used to transform *E.coli*. The transformants were plated on kanamycin plates to determine the number of transformants obtained or kanamycin plates containing different concentrations of moxalactam (mox) to obtain moxalactam resistant clones. The mox-resistant clones were further characterized to determine the fold increase in resistance compared to cells containing the wild type ampC gene. Ten mox-resistant clones were obtained, which had a fold increase in mox-resistance ranging from around 3-fold to 20-fold (0.8-6 μg/mL) above wild type (0.3 μg/mL).

Sequencing of the *ampC* gene in the plasmids from these variants revealed that each of them contained one to three of the selected library amino acid changes in ampC (Table 4). Two of the variants, IRL1.8.4 and IRL1.8.5 also contained additional mutations introduced during the PCR process (Table 4). The IRL1.6.1 variant, which has a 20-fold increase in mox-resistance was the best variant in this library and had two changes at positions S288 and R348. The substitutions Y220N, A219P and L61M appeared in more than one clone suggesting that they may be important for conferring resistance. Thus, this example shows that the application of a distance-based constraint onto a scoring matrix was successful in producing ampC variants that had a significantly higher resistance to the antibiotic moxalactam.

Example 3. Alteration of β-lactamase Specificity Using a Recruitment Matrix

This Example demonstrates the application of a distance-based constraint vector to the *E. cloacae* ampC molecule and recruitment of amino acids observed in other ampC proteins.

To create the IRL library in this example, first, the sequence of the ampC protein from *E. cloacae* (reference sequence) was aligned with ampC protein sequences from *A. sobria*, *E. coli*, *O. anthropi*, *P. aeroginosa*, *S. enteriditis* and *Y. enterolitica* using the AlignX program from Vector NTI Suite (Informax Inc. Bethesda, MD). Those positions in the alignment where amino acids other than those found in the reference sequence were observed were recruited, and a distance-based constraint vector was applied to these positions to limit mutations to residues that

were surface exposed and 6 angstroms from the binding site of ligands to the *E. cloacae* ampC 3-D structure. Specifically, the *E. cloacae* ampC crystal structure (Protein Database Base ID# 1BLS) and 6 *E. coli* ampC structures containing bound inhibitors or substrates (Protein Database Base structures 1C3B, 1FCM, 1FCN, 1FCO, 1FSW, 1FSY) were first loaded into the program MOE 2000.01 (Chemical Computing Group, Inc., Montreal Canada). Because each structure consists of a homodimer, one of the monomers and its associated ligand was deleted. Next, the main chains of all the structures containing bound ligands were aligned (0.4 angstroms RMS deviation) and all the water molecules were manually deleted. The main chains of all structures except the *E. cloacae* structure (1BLS) were then removed. The resulting structure consisted of the *E. cloacae* ampC molecule with all of the superimposed ligands from the other 6 ampC structures. All surface-exposed side chains (i.e. did not count the backbone, just the beta carbon, and outward atoms) in ampC with atoms within 6 angstroms of the ligand atoms were then selected for the IRL library. Eight positions were selected and substitutions were chosen based on the amino acids observed at those positions in other members of the ampC protein family used in the alignment. This library was termed the 'recruitment library' or IRL2 library.

To create the IRL2 DNA library, 15 mutagenic forward primers containing the different substitutions were designed and used in a PCR reaction containing a single wild type reverse primer and the *E. cloacae ampC*-containing plasmid pAL20 as template. After digestion of the methylated template DNA using the *DpnI* enzyme, the unmethylated PCR product was used to transform *E.coli*. The transformants were plated on kanamycin plates to determine the number of transfomants obtained or kanamycin plates containing different concentrations of moxalactam (mox) to obtain moxalactam resistant clones. The mox-resistant clones were further characterized to determine the fold increase in resistance compared to cells containing the wild type ampC gene. Fifteen mox-resistant clones were obtained, which had a fold increase in mox-resistance ranging from around 3 fold to 83 fold (0.8-25 μg/mL) above wild type (0.3 μg/mL) in a single round.

Sequencing of the ampC gene in the plasmids from these variants revealed that 12 variants contained one to three of the desired library amino acid changes in ampC (Table 4). In addition to the desired mutations observed in the winners, some of the winners had additional unexpected mutations which may have contributed to the phenotype in some cases. Four of the variants contained additional unexpected mutations either in the promoter or within the ampC

gene due to errors in the PCR process. These included S263P in IRL1.8.4, S17T in the signal sequence in IRL1.8.5, A217V in IRL2.8.4, and T125M in IRL2.3.6. The observation that 3 of the 15 variants contained wild type ampC sequence suggests that mutations elsewhere in the plasmid vector or in the E. coli genome can contribute to the phenotype, which is not unexpected. Silent muations were also seen at position A351 in IRL1.8.10, S286 in IRL2.8.3, and at A152 in IRL2.8.14. Promoter region mutations were seen in IRL2.8.7 (a to g at +168), IRL2.8.12 (c to t at +136), and IRL2.8.13 (c to t at +237 and t to c at +205).

The substitutions V120F and N345I appeared in several clones suggesting their importance for increasing mox resistance. Although it can be argued that these mutations came up several times due to PCR primer bias, the sequencing of random library clones not selected for mox resistance did reveal other positions where a large number of substitutions were seen, but which did not show up in the variants. It is interesting that compared to the IRL1 library, the IRL2 library shows a different profile of substitutions in the variants. Again, this example shows that the use of a distance-based constraint and recruited residues from multiple sequence alignment were successful in producing ampC variants that had a significantly higher resistance to the antibiotic moxalactam.

Molecular Biological Methods

The mutagenic primers used for creating the PCR-based DNA libraries each contained 37 bases with 17 bases flanking the mutant codon on both sides. All mutagenic and wt primers used for creating the DNA libraries or for sequencing were obtained from Operon Technologies (Alameda, CA).

A single reverse primer and 90 IRL1 or 15 IRL2 mutagenic forward primers were used in a PCR reaction with a template, plasmid pAL20 containing the *E. cloacae ampC* gene. Plasmid pAL20 was created by sub-cloning the *ampC* gene into the TOPOBLUNT vector (*kan*^{r}) obtained from Invitrogen (Carlsbad, CA). The final reaction contained 0.5μM of the reverse primer and 0.5μM of all IRL forward primers combined (all primers together were 25 pmols), 16 fmol of pAL20, 15 nmol of each dNTPs, 5 units of the Herculase polymerase (Stratagene, La Jolla, CA) and a Herculase-specific buffer also from Stratagene. The total reaction volume was 100μL. The cycling conditions included an initial cycle at 94°C for 3 minutes followed by 30 cycles each containing a step at 94°C for 30 seconds, a 55°C step for 30s and a 68°C step for 5

minutes. A final elongation cycle at 68°C for 7 minutes was also included. An MJ Research PTC thermal cycler was used for the PCR reaction. After the PCR reaction was carried out, the plasmid template in each of the PCR reactions was digested with the *DpnI* enzyme, which cleaves the methylated DNA template and not the PCR product.

For each library, 1μL of the *DpnI* digested PCR reaction was transformed by electroporation into TOP10 one-shot electrocompetent cells from Invitrogen. The electroporation was conducted using a BIORAD electroporator. A fifth of the transformation mix was plated on LB plates containing 50μg/mL kanamycin (kan) and the remaining mix was plated on LB plates containing 50μg/mL kan and 0.5μg/mL moxalactam (mox; obtained from Sigma). Between 2000 and 4000 transformants were obtained per transformation based on the number of colonies observed on the kan plates. Several transformations were carried out to obtain 21000 and 54000 colonies for the IRL1 and IRL2 libraries respectively. Those transformants that grew on plates containing mox were streaked for single colonies on LB plates containing 50μg/mL kan and 0.5μg/mL mox. A single colony from each of the mox-resistant clones was used to inoculate 200μL of LB containing kan in a 96 well microtiter plate. The plate was grown at 37°C with shaking for 18 hours, and each of the cultures in the wells was diluted 10,000-fold into 12 microtiter plates containing LB with different concentrations of mox (0 to 100μg/mL). Kanamycin was also added to the media to maintain selection for the *ampC* pAL20 plasmid. After incubation at 37°C with shaking for up to 21 hours, the absorbance of the cells grown in each well was measured at 600nm. The fold increase in mox resistance was calculated based on the extent of growth of cells containing the wild type *ampC* gene. Plasmids were extracted for sequencing from all library clones that had a mox resistance of greater than 2.5 fold compared to wild type.

Example 4. Generation of a Conservation Index as a Constraint Vector

A conservation index may be defined as a measure of the degree of conservation at each position in a multiple sequence alignment. A conservation index algorithm developed by Novere et al. (Biophys. Journal v.76 , p. 2329-2345, May 1999) was used to generate a conservation index based on the alignment of the ampC proteins. A conservation index was assigned at each position in the alignment according to the equation:

$$CI = \frac{\sum_{i=1}^{N} \sum_{j>i}^{N} S_{ij}}{\sum_{i=1}^{N} \sum_{j>i}^{N} S}$$

where N is the number of sequences in the alignment, S_{ij} are the global similarities of the ith and jth sequences, and s_{ij} is the relevant similarity matrix element for the sequences i and j at the given position. The default similarity matrix from the Wisconsin package program GAP (Devereux et al., 1984) can be used, rescaled to [0-100]. The resulting values range from 0 to 100. A score of 100 indicates absolute conservation.

Although the invention has been described in some detail with reference to the preferred embodiments, those of skill in the art will realize, in light of the teachings herein, that certain changes and modifications can be made without departing from the spirit and scope of the invention. Accordingly, the invention is limited only by the claims.